Friedrich Schiller University Jena

Faculty of Social and Behavioral Sciences

Department of Psychology

Hans Is Clever After All – Large Number Discrimination and Intuitive Statistics in Domestic Horses (*Equus caballus*)

Submitted by Melanie Henschel

Matriculation number: 138010

Bachelor thesis to obtain the academic degree Bachelor of Science (B.Sc.)

Primary supervisor:

Prof. Dr. Daniel Haun, Leipzig University

Secondary supervisor:

PD Dr. Matthias Reitzle, Friedrich Schiller University Jena

Jena, 06.09.2016

Table of contents

Abstract

The current study should, first, answer the outstanding question whether domestic horses can discriminate not only small but large discrete quantities as well, using a discrimination task between two different-sized sets of food items. Second, this study examined basic statistical reasoning in horses. Subjects had to choose between two one-item samples drawn from two visible distributions of stones and food items. In both tests horses showed a mean success rate significantly above chance level. A third experiment controlled for use of olfactory cues where mean success rate was at chance level, thus excluding this explanation. Therefore, this study proves large number discrimination in horses and is a first indicator that the ability to make inferences about single-event probabilities is not exclusive to higher primates.

Abstract in German (Zusammenfassung auf Deutsch)

Die aktuelle Studie soll erstens die ausstehende Frage beantworten, ob Hauspferde in der Lage sind nicht nur kleine, sondern auch große Mengen zu diskriminieren, wofür eine Diskriminationsaufgabe zwischen zwei unterschiedlich großen Mengen von Futterstücken durchgeführt wurde. Zweitens untersuchte diese Studie grundlegendes statistisches Schlussfolgern bei Pferden. Die Versuchstiere mussten zwischen zwei Stichproben wählen, je bestehend aus einem Objekt, die aus zwei sichtbaren Verteilungen von Steinen und Futterstücken gezogen wurden. In beiden Tests zeigten die Pferde eine signifikant über dem Zufall liegende Durchschnittserfolgsrate. Ein drittes Experiment diente als Kontrolle für olfaktorische Hinweisreize, wo die durchschnittliche Erfolgsrate nicht signifikant über dem Zufall lag, daher kann diese Erklärung ausgeschlossen werden. Diese Studie beweist demzufolge Diskrimination großer Mengen bei Pferden und ist ein erster Hinweis, dass die Fähigkeit, Schlussfolgerungen über Wahrscheinlichkeiten einzelner Ereignisse zu ziehen nicht ausschließlich bei höheren Primaten vorliegt.

Hans Is Clever After All – Large Number Discrimination and Intuitive Statistics in Domestic Horses (*Equus caballus*)

If one is looking for studies about cognitive abilities of horses, one is bound to constantly stumble across studies mentioning the so-called *Clever Hans effect* because this effect was named after a horse. He got the name *Clever* Hans because he could allegedly solve complicated mathematical tasks. But after countless evaluations of the case judges finally came to the conclusion that Hans read subtle, unintended cues given by his human opponent which told him the solution for the task. There was vast disappointment about Hans obviously not being as clever as everybody had thought. However, the cues that Hans read were apparently even too subtle for most humans (Gross, 2014).

Probably, Hans' story was not entirely uninvolved in cognitive science not paying a lot of attention to horses in the past, leaving the actual interesting question unanswered: How "clever" was Hans really? Past studies mostly addressed basic cognitive abilities such as conditioning and generalization, in some cases also categorization or concept learning, in all of which horses performed very well (for an overview see Hanggi, 2005). Today, at least we know as much: Although Hans did not solve the tasks he was given the way he was supposed to, if he would have been asked in the right way he would in fact have been able to count to three and solve simple addition problems. Uller and Lewis (2009) could show that domestic horses can discriminate sets of up to three objects, even if the objects were put into opaque buckets one by one. To solve this problem, the animals have to be able to memorize how many objects are already in the bucket and add the following internally. This version of numerical discrimination requires a lot of mental capacities and is challenging even for apes (Hanus & Call, 2007). A more recent study could further support these findings, proving horses' numerical discrimination abilities using a new methodological paradigm (Petrazzini, 2014).

Still, the current amount of studies about horses remains relatively small but slowly, more and more evidence of advanced cognitive abilities in horses is piling up. There are many indications that they are capable of far more complex cognitive processes than it is generally assumed.

But what actually makes a creature intelligent or clever for that matter? According to behavioral ecologists intelligence, like every other feature of a living being, is formed through evolutionary processes and therefore specific cognitive abilities represent adaptations to specific environmental challenges (Roth & Dicke, 2005). Species considered as intelligent thus have a qualitatively or quantitatively greater aggregate of those adaptations.

The *social intelligence hypothesis* proposes that high cognitive abilities evolved because of the challenges of a complex social life (Humphrey, 1976). This hypothesis agrees with other theories in that social contexts generate especially complex adaptive pressure towards organisms. Horses live in such complex social structures: They maintain a strict hierarchy within their herd (Houpt et al., 1978), form stable friendships (Feh & De Mazières, 1993 and Sigurjónsdóttir et al., 2002) and they even infer information about their own position within the group from the observation of social interaction between others (Krüger & Heinze, 2007), to name just a few findings. Therefore, we can assume that horses have a noteworthy set of cognitive abilities. Additionally, it is probable that this set of cognitive abilities has a lot in common with our very own species since a lot of our intelligence evolved through adaptation to social environments as well.

One cognitive faculty we long thought to be uniquely human is statistical reasoning since a lot of studies claimed that this ability develops late in ontogeny (Piaget & Inhelder, 1975) and even then stays unstable (Tversky & Kahneman, 1974, 1981). But a new body of research found intuitive statistical abilities already to be present in human infants

(Denison & Xu, 2010 and Xu & Garcia, 2008). Building on that, Rakoczy and his team (2014) were the first to expand this research to non-human animals. They could proof that great apes (chimpanzees, gorillas, bonobos and orangutans) are able to solve the simplest form of statistical problems: comparing relative frequencies, "that is, frequencies of items of a given kind in a population (say winner tickets in a lottery) relative to the frequencies of all kinds of items in the population (all tickets)" (Rakoczy et al., 2014) and drawing flexible inferences from these populations to samples and vice versa. Rakoczy et al.'s results show that intuitive statistics most likely evolved before the evolutionary separation of humans and apes. Given the huge advantage this ability provides in a variety of contexts, it seems probable that its heritage roots even deeper in evolutionary history, maybe dating back 80 to 90 million years ago, when horses shared their last common ancestor with primates (Eizirik et al., 2001) or even further.

The ability to discriminate absolute quantities seems to be a necessary but not a sufficient precondition for comparing relative frequencies, since it can be assumed that the ratio within a population corresponds to a comparison between two absolute quantities. Both abilities ought to be intuitive and therefore present without prior training to form associations about reinforcement contingencies.

Quantity processing has been found to work through two core systems of number. The study by Uller & Lewis (2009) and related articles (e.g. Agrillo et al., 2012) support the idea of an *object file system* with which up to three or four objects (depending on species and age) can be computed precisely and simultaneously; this ability is also called *subitizing* (Kaufman et al., 1949). Other studies (see below) showed that especially sets of more than four objects are computed with another system, the *analog magnitude system* with which (larger) quantities are estimated approximately. This system is ratio dependent (Weber's Law), i.e. if an individual is able to discriminate 6 versus 12 it can also

discriminate 24 versus 48 and it will show worse performance comparing 10 to 12 etc. (Feigenson et al., 2004).

To my knowledge, discrimination of large sets of objects, and therefore the existence of an analogue magnitude system, has not been tested explicitly in horses. However, every species studied so far was able to pass this test and showed ratio dependency (referring to published articles). The spectrum of species also shows a great variety from amphibians (Krusche et al., 2010) to fishes (Agrillo et al., 2010) and birds (Rugani et al., 2014) to mammals (Vonk & Beran, 2012). In combination with the studies showing numerical discrimination ability of small sets in horses (Petrazzini, 2014 and Uller & Lewis, 2009) it seems likely that horses possess equal quantity processing.

Therefore, I examined if horses are capable of discriminating large sets of objects in Experiment 1 (Exp. 1), using a classic discrimination task with sets of food items. Furthermore, I conducted a modified version of Racoczy et al.'s procedure (2014) with domestic horses. In this procedure subjects are presented with two visible populations which have different distributions of objects of two kinds, one of which is preferred over the other. The experimenter then draws a one-item sample from each population with the kind of item remaining invisible and makes them available for the subject to make a choice. This procedure tests whether subjects can discriminate between the two populations using information about the ratios of each kind of item and furthermore it requires subjects to form expectations about the probability of each sample to contain the preferred object. Experiment 2 (Exp. 2) of this study thus tested if horses are capable of intuitive statistics in the same manner as apes are.

Experiment 3 (Exp. 3) ruled out olfactory cues as alternative explanation.

I hypothesized that horses would perform significantly above chance level in Exp. 1 and Exp. 2 but at chance level in Exp. 3, meaning domestic horses are able to discriminate

large sets of objects, are capable of basic statistical inferences and do not rely on olfactory cues to solve these tasks.

Thereby, this study, first, answer the outstanding question whether horses share the ability to process large quantities with many other species, second, will give us further indications towards the evolutionary age and heritage of intuitive statistics and, third, will give us new information about the cognitive abilities of domestic horses and maybe repair a bit of the damage Clever Hans has done to their reputation.

Subjects

Participants were recruited from a group of 34 adult and three adolescent domestic horses (*Equus caballus*) housed at the "Paulushof" in Zwickau. The horses are kept in boxes (mostly individually, some with up to four others) with daily access to a paddock within the group. Seven horses did not pass the training due to lack of motivation and one was excluded from all further testing after completing training because it was afraid of the buckets. This lead to a final sample of 29 adult horses (17 mares, 7 stallions and 5 geldings), consisting of four different breeds: Deutsches Sportpferd ($N = 11$), Shetland pony ($N = 15$), Sachsen-Anhaltiner ($N = 1$) and Schweres Warmblut ($N = 2$). Three of the mares were pregnant at the time of data collection. Age ranged from 3 to 25 years with a mean age of 11.66 years. The sample comprised trained horses used for lessons ($N = 5$), leisure horses ($N = 3$), breeding horses ($N = 12$), young horses in training ($N = 8$) and one horse put out to grass ($N = 1$). None of the subjects had prior experience with cognitive studies. (For more detailed information about each subject see Appendix A.)

All subjects of this sample participated in all three experiments.

Materials and general procedure

Subjects were confronted with pieces of food and stones of approximately the same size (∅ 1.5 to 3 cm; height 0.5 to 1.5 cm) and shape but different color. All horses

were first trained with carrots but if the subject did not take the reward it was switched to something they liked. Consequently, three horses received pieces of apple for the rest of the procedure. In contrast to Rakoczy et al. (2014) I decided not to use two different kinds of food, first, because horses are usually less selective with their food and, second, because they are less familiar with the testing situation than the apes. Discriminating between food and non-food should keep learning effort at a minimum and make the task easier for the horses. Additionally, this also rules out a change in food preference during the test due to repletion or daily mood as it can be assumed that horses always prefer food over non-food.

Except for the first pretest, items were presented in transparent buckets (∅ 22.5 cm; height 19.5 cm) closed with transparent lids. For the second pretest, Exp. 2 and 3 the lids had holes (∅ 10 cm) cut into them which allowed sampling but at the same time prevented the horses from accessing the food if the bucket got into their reach.

Subjects always had to choose between two options. The one they touched first with their snout was counted as chosen. If subjects chose food, they immediately received it. If they chose a stone, it was moved out of the subjects' reach so they could not accidently eat it and presented to them. The side with the favorable choice was counterbalanced across trials to assure that a simple preference for one side would not lead to success above chance level. If a subject did not choose, the experimenter moved the options back out of the subject's reach and then forth again. The same was done if a subject touched both sides simultaneously.

Horses were trained and tested either in their individual boxes ($N = 17$; see Fig. 1) or at the place they are usually groomed ($N = 12$; see Fig. 4). Both places provided visual and/or acoustic and olfactory access to the other horses so that subjects would be calm and able to

concentrate on the task. All horses were tied during the procedure.

To control for Clever Hans effect the experimenter wore a plain grey cardboard mask (DIN-A4) with two small slits for vision (0.5 x 3 cm) during the whole training and testing procedure.

To prevent subjects from smelling the location of the food item, rubber gloves were worn during training and testing. At the beginning of each session the subject could smell the gloved hands, was stroked and could smell them again. This procedure should assure that subjects were not afraid of touching the gloved hands and were therefore willing to indicate their choices.

There was at least one day between the last session of training and the first test session for each subject. The same applied to all experiments, i.e. only one session consisting of 12 trials was conducted with each subject per day.

With the beginning of testing, subjects and experimenter were filmed. Subjects' performance was scored live and later compared with the recordings. The order of the experiments was randomized and counterbalanced across subjects to prevent effects of order or position.

Training

First, a pretest was conducted to make the horses familiar with the basic principle of the study and to rule out subjects that were not motivated to participate. The experimenter held one item of each kind in hand and presented it to the subjects. Hands were both closed as soon as the experimenter was sure subjects were watching and they were then given the choice between the two options. Stimuli were drawn from opaque buckets.

Subjects passed the pretest when they chose the hand with the food in 10 out of each last 12 trials. 30 horses passed the initial training. Required trials ranged from 12 to 138.

Because of this large variance I conducted a second pretest in which transparent buckets were used, one containing pieces of food and one an equal amount of stones. This procedure should assure that subjects would focus on the content of the buckets during the test. First, one bucket at a time was shown to subjects from all sides. Then one item was visibly drawn from the bucket. Subjects could see, smell and taste each kind of item and eat the piece of food. The experimenter then presented both buckets simultaneously, positioned them next to each other in front of the subject and drew one item of each bucket with hands remaining closed until the subject had chosen.

The cut-off remained the same. All 30 horses also passed the second pretest with required trials ranging from 12 to 51 (for more detailed information about training performance see Appendix B).

Recollection trials

At the beginning of each test block the procedure of the second pretest was repeated until subjects chose correctly three consecutive times. This should assure that subjects remembered the procedure and were ready to be tested.

Experiment 1: large number discrimination

Design and procedure

One bucket contained 5 and one 10 pieces of food (ratio 1:2). Buckets were closed with transparent lids (without holes), then shown to subjects from all sides one at a time. Both buckets were then held next to each other in front of the subjects for several seconds and then moved within their reach for them to make a choice. As soon as subjects had chosen, they received the content of the chosen bucket as reward. The procedure is depicted in Fig. 1.

Figure 1. *In Exp. 1 subjects are first presented with the two buckets (A) and are then given the choice between them (B). Horses indicate their choice with their snout (C) and immediately receive the content of the chosen bucket (D).*

Statistical analysis

All analyses were done with R software (version 3.3.1). First, a one-sample t-test against chance (50%) was conducted to test whether subjects' overall performance (mean percentage correct) was better than expected by chance. First-trial performance was assessed with a one-tailed binomial test. Improvement across the course of the experiment was assessed with a Pearson's correlation between trial number and sum of correct choices per trial. It was also tested for effects of gender on performance (one-way ANOVA for three genders, unpaired two-sample t-test for two), in case an ANOVA revealed significant differences, a Tukey's test assessed their exact location. A Pearson's correlation tested for effects of age and an unpaired two-sample t-test for effects of breed on test performance. Regarding the latter, only subjects of the breeds Deutsches

Sportpferd and Shetland pony were included because for the other two breeds the groups were too small.

Results

As can be seen in Fig. 2, 75% of horses chose correctly in more than half of the trials. Horses as a group chose the bucket with 10 pieces of food in 65% of the trials (Fig. 3), significantly above chance level ($SD = 0.15$, $t[28] = 5.36$, $p < .001$, Cohen's $d = 1.00$). This pattern was also reflected in first-trial performance where 23 (79%) of the horses chose the bucket with the larger amount of food items, significantly more often than expected by chance ($p = .001$). Additionally, no significant improvement of performance over the course of the experiment could be detected ($r[10] = -.358$, $p = .253$). (For detailed information about test performance see Appendix C.) No significant differences in performance due to gender, age or breed were found (see Tab. 1).

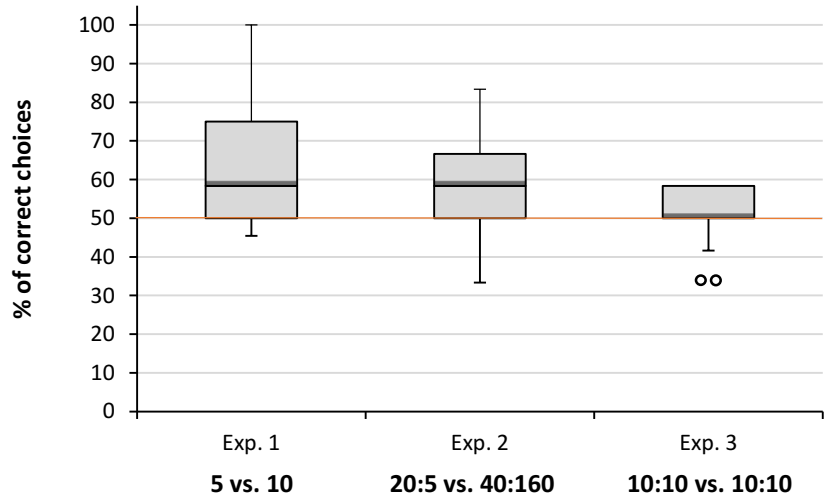

Figure 2. *Percentage of correct choices in Exp. 1, Exp. 2 and Exp. 3. Box plots show first (Q_1) and third quartiles (Q_3), medians and outliers (dots). The end of the upper whisker marks the last value within Q_3 + 1.5 * IQR (interquartile range), the end of the lower whisker marks the last value within Q_1 – 1.5 * IQR. The orange line represents chance level.*

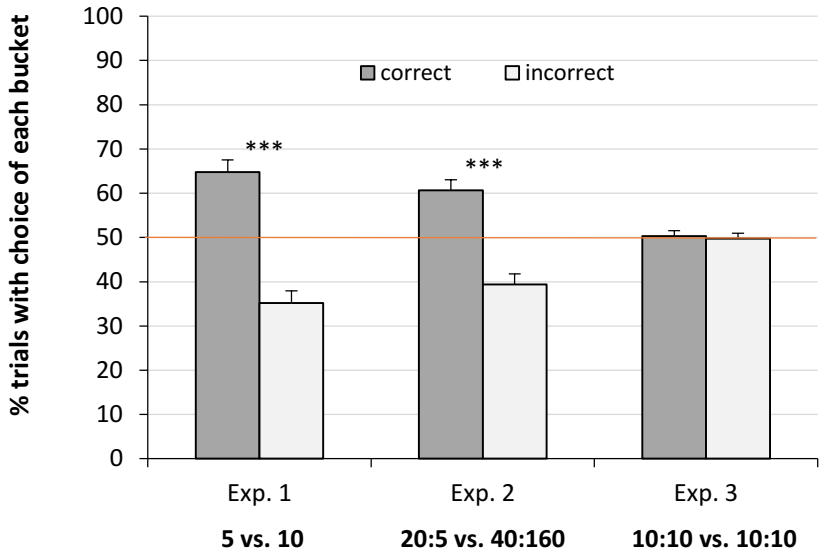

Figure 3. *Mean percentage of trials (with standard errors) in which subjects chose the correct/incorrect bucket for Exp. 1, Exp. 2 and Exp. 3. The orange line represents chance level.*

Table 1. *Test statistics for Exp. 1*

Variable (with values)	**df**	**Test statistic**	**Significance**	**Effect size**
gender (females, males)	27	t = -1.00	.328	Cohen's d = -0.38
gender	2	F = 0.55	.583	η^2 = 0.04
(mares, geldings, stallions)	26			
age	27		.711	r = -.072
breed	24	t = 1.35	.189	Cohen's d = 0.54

Discussion

Results confirmed that horses are able to discriminate sets of food items also within the large number range. Furthermore, there was no significant difference between three or two genders, but between males and females the effect was of small to medium size, indicating that males performed somewhat better than females. Similarly, the

difference between breeds in performance was not significant but of medium size, suggesting that horses of the breed Deutsches Sportpferd performed moderately better than Shetland ponies. Both effects could be valid and only non-significant for lack of power. The effect of age was neither significant, nor of any noteworthy size.

Since the general choice pattern was also mirrored in trial 1 and there was no significant improvement in performance over the course of the experiment, the findings cannot be attributed to learning over trials.

On the other hand, the aforesaid correlation, though non-significant, was of medium size and negative, indicating that horses' performance declined over trials. During the procedure it could be observed that many subjects consistently chose the bucket with more food items until they, maybe accidently or out of curiosity, chose the other bucket for the first time. After that, they did not seem to care anymore which bucket they chose. This behavior could indicate that the difference between receiving 10 or 5 pieces of food just was not a big enough motivation to get the horses thinking about their choices. All subjects are well-fed and receive carrots or apples regularly. In addition, it is possible that on the basis of the training, the premise seemed to be that they would not receive any food if they chose the wrong bucket. Since they did get rewarded every time, it is possible that it did not seem to matter anymore which bucket they chose. This could explain the relatively small overall performance of only 65%. In trial 1 performance was as high as 79%, suggesting that horses' numerical discrimination ability is better than the current findings imply. Future studies should focus on this problem and also use other paradigms to examine equine number sense.

Since the current study only examined one ratio between the two buckets, it cannot be said whether horses have an approximate magnitude system as proposed by Feigenson et al. (2004). Future studies should vary ratios to find out whether horses' numerical

discrimination ability follows Weber's law like in other species (e.g. Agrillo et al., 2012 and Krusche et al., 2010).

Experiment 2: statistical inferences

Design and procedure

Exp. 2 tested for statistical abilities and was modelled after Rakoczy et al. (2014) with a few adjustments. Subjects were confronted with two distributions of food items and stones. The first bucket contained 20 pieces of food and 5 stones (ratio 4:1), the second contained 40 pieces of food and 160 stones (ratio 1:4). The buckets were first shown to subjects from all sides one at a time, then presented next to each other for several seconds before they were put on the ground and a one-item sample (always of the majority kind) was drawn from each population in the same manner as in the training. Subjects then made their choice and were rewarded accordingly (see Fig. 4).

Figure 4. *In Exp. 2 subjects are first presented with the two buckets (A). The experimenter then draws a sample from each bucket (B) and subjects are given the choice between them (C). Horses indicate their choice with their snout (D) and are shown the stone (E) or receive the food (F).*

Results

Of all horses, 75 % chose correctly in more than 50% of the trials (Fig. 2). Horses as a group chose the sample from the bucket with the better distribution in 61% of the trials (see Fig. 3), significantly more often than expected by chance ($SD = 0.13$, $t[28] = 4.42$, $p < .001$, Cohen's $d = 0.82$). In trial 1, horses failed to replicate this choice pattern, only 18 (62%) of the subjects chose the sample from the correct bucket ($p = .133$). However, no significant improvement of test performance over trials could be detected ($r[10] = .321$, $p = .309$). No significant effects on test performance of age or breed could be found but there was a significant effect of gender. Post-hoc analysis revealed a significantly higher success rate of mares compared to stallions but there were no significant effects of any other constellation (see Tab. 2).

Table 2. *Test statistics for Exp. 2*

Variable (with values)	**df**	**Test statistic**	**Significance**	**Effect size**
gender (females, males)	27	$t = 2.78$	.010	Cohen's $d = 1.05$
gender	2	$F = 4.71$	.018	$\eta^2 = 0.27$
(mares, geldings, stallions)	26			
mares – geldings[a]			.433	
stallions – geldings[a]			.444	
stallions – mares[a]			.015	
age	27		.303	$r = .198$
breed	24	$t = 0.53$	.603	Cohen's $d = 0.21$

[a] Tukey pairwise comparisons

Discussion

Results confirmed the hypothesis that domestic horses are able to discriminate relative frequencies and to draw inferences from populations to samples and therefore to make predictions about single-event probabilities. Furthermore, the findings show a difference in performance between genders, i.e. mares performing better than stallions but since the number of stallions and geldings was quite small, it is not unlikely that these findings are artefacts and of no substantial meaning. Still the effect is of large size and sex differences in horse cognition have been reported before (Murphy et. al, 2004 and Wolff & Hausberger, 1996). Regarding age, there seems to be a small, although non-significant effect on performance, in that older horses performed slightly better than younger ones. Similarly, there is a small but non-significant effect of breed, indicating that subjects of the breed Deutsches Sportpferd performed slightly better than Shetland ponies.

But then, do the present findings actually reveal statistical faculties or are there simpler explanations? The design of this experiment excludes the possibility to only rely on absolute frequencies for discrimination, since there is a conflict of absolute and relative frequency, i.e. the bucket with the greater number of food items is the one with the worse distribution and therefore delivers stones and vice versa. In addition, I used 5 instead of 3 as smallest number (in contrast to Rakoczy et al., 2014) because as has been proven horses can subitize up to three objects (Uller & Lewis, 2009), therefore it seemed possible that horses might count how many objects of that kind (in this case stones) were still left in the bucket and make a conclusion from that as to which kind of item must be hidden in the experimenter's hand. By using 5 objects this could be prevented. Thus, horses were forced to base their decision on relative rather than absolute frequencies.

However, there are indications that simpler choice heuristics were used, in that subjects learned which of the buckets is the one that delivers food and chose accordingly. The fact

that subjects did not choose correctly significantly above chance level in trial 1 supports this explanation. Moreover, though there was no significant improvement of performance over the course of the experiment, this effect was still of medium size, meaning that there was a moderate improvement over trials that might just not have become significant because of missing power. On the other hand, it does not seem very likely that horses learned to discriminate between the two buckets so quickly, given the large number of trials required for the initial training in which the discrimination task was much easier. But there is also another possible explanation: Horses, in contrast to humans, do not have cones for long-wave light (red) in their retinas. Consequently, the mainly used carrots look greenish or yellowish to horses. The stones used for this study had a white-yellowish color. Thus, it could be possible that, at least for some subjects, it was harder to discriminate the two kinds of objects than planned, especially when they were mixed together in one bucket. Horses would only have learned to discriminate the two kinds of objects accurately over trials, hence the improvement over time.

All things considered, the current findings are a first indicator that statistical abilities might not be exclusive to primates but at this point it remains unclear if horses are really capable of using statistical information to solve the given task. More research needs to be done, varying more different distributions and only conducting one trial per session to prevent learning effects from the start. By doing so, we can also determine if Weber's law also applies to statistics, i.e. if the discrimination of relative frequencies is a function of the ratio of these frequencies to each other.

Experiment 3: control for use of olfactory cues

Design and procedure

Exp. 3 served as control condition to rule out that horses would smell which hand contained the food item despite the rubber gloves. To do so, the buckets both contained a

distribution of 10 food items and 10 stones. It was counterbalanced across trials on which side the correct choice was and from which bucket it was drawn. In contrast to Rakoczy et al. (2014), I did not use opaque buckets because it seemed possible that this would disturb the symmetry to the experimental condition but with an even distribution of food and stones it was still not possible to determine the correct choice at sight. Presentation and sampling procedure were identical to Exp. 2.

Results

Only half of the horses chose correctly in more than 50% of the trials with a maximum of 58% correct choices (see Fig. 2). Horses as a group chose the hand with the food item in 50% of the trials (see Fig. 3), not significantly different from chance (SD = 0.07, t[28] = 0.23, p = .412, Cohen's d = 0.04). This pattern was already present in trial 1, where only 16 (55%) of horses chose correctly (p = .356) and no significant improvement over trials could be detected either (r[10] = -.196, p = .542). No significant differences due to age or breed could be detected. If females were compared to males, no significant differences could be found either but if males were split into geldings and stallions, a significant effect of gender emerged. Geldings had significantly higher success rates than mares and stallions but no significant difference between mares and stallions could be detected (see Tab. 3). An additional analysis was conducted which revealed that the effect mentioned above is most likely due to a confounding of gender with the purpose the horses were used for. Trained horses used for lessons performed significantly better than horses used for another purpose (unpaired two-sample t-test, t[27] = -2.49, p = .019, Cohen's d = -1.22). (For results from analyses adjusted for outliers see Appendix D.)

Table 3. *Test statistics for Exp. 3*

Variable (with values)	df	Test statistic	Significance	Effect size
gender (females, males)	27	t = -1.19	. 243	Cohen's d = -0.45
gender	2	F = 5.69	. 009	η^2 = 0.30
(mares, geldings, stallions)	26			
mares – geldings[a]			.013	
stallions – geldings[a]			.013	
stallions – mares[a]			.860	
age	27		. 121	r = .295
breed	24	t = 1.18	. 251	Cohen's d = 0.47

[a] Tukey pairwise comparisons

Discussion

Horses did not detect the location of the food by use of olfactory cues. Analyses revealed that geldings performed significantly better in this experiment than any other gender. This was especially surprising since in this experiment all horses were expected to have approximately the same success rate (50%). In fact, all geldings showed exactly the same performance, i.e. 7 out of 12 successes (58%) which is exactly one more correct choice than 50%. Looking at the behavior of those horses, there are indications that they tried harder than other horses to solve the pattern with which the experimenter sampled, which led to a slightly higher success rate. This led to the consideration that this effect might have nothing to do with gender but how used to solving tasks the concerned horses are. As a matter of fact, except for one subject geldings and trained horses were the same animals. The analysis revealed that in fact trained horses performed better than other horses, so the above-mentioned effect is unlikely a matter of gender.

Age of subjects had only a small to medium, though non-significant influence on performance, in that older horses performed moderately better. Additionally, subjects of the breed Deutsches Sportpferd again performed moderately, though not significantly better than Shetland ponies.

The choice pattern at chance level was also mirrored in trial 1 and performance did not improve over the course of the experiment either, proving that the task gave away no information based on which horses could learn to discriminate the two buckets, only leaving the smell as source of information.

General discussion

Numerical cognition was long thought to be uniquely human. In the last years a big body of research could show that not only humans have this ability but that it is wide-spread over the animal kingdom. Horses have already been proven to be able of discriminating continuous quantities (Hanggi, 2003), as well as small discrete quantities (Petrazzini, 2014 and Uller & Lewis, 2009). This study filled the long existing gap in this body of research, finally proving that domestic horses are also capable of discriminating large discrete quantities. Furthermore, this study is the first to have investigated statistical abilities in equines, an even more complex set of mathematical faculties that was long thought to be acquired late in ontogeny even in humans (Piaget & Inhelder, 1975). The present findings now deliver first indications that the distantly related primates and equines might not only share abilities for processing absolute but also relative frequencies which is a fundamental basis for statistical reasoning.

But what do these findings imply for comparative psychology? First, the present study further supports the *core knowledge hypothesis* according to which mental representations of quantity (among others) are innate and shared at least among vertebrates (Spelke & Kinzler, 2007), maybe even invertebrates as well (Gross et al., 2009).

Second, the present findings are a first step for answering one of Tinbergen's four questions: How common among animals is the behavior indicating statistical reasoning and thus the underlying cognitive ability (phylogeny; Tinbergen, 1963)? The findings of Racoczy et al. (2014) were the first to show that statistical abilities evolutionarily predate humans. Based on the current findings, the most likely implication is that statistics either root even deeper in evolution or developed several times and independently in different species, i.e. through convergent evolution.

Given that hominids, i.e. humans and great apes, as well as horses live in complex social groups it is possible that intuitive statistics are limited to highly social species because their mode of life gave them the tools to solve complex cognitive problems, also outside of social contexts (Humphrey, 1976). At this point, we do not know yet if statistics evolved because of social causes but arguments in favor of this explanation come to mind. Social contexts require an individual to calculate the likeliness of behavioral responses of others to their own behavior and the balance of advantage and disadvantage (Humphrey, 1976), all of which can be put in terms of probabilities. For example, a train of thought in a social interaction might be: "If I steal my counterpart's food, how likely is it that he/she will be aggressive towards me and how likely is it that he/she will tolerate it?" Of course such problems can be solved by trial and error but being able to reason about probabilities without previous knowledge about frequencies of events bears an obvious advantage, especially in social contexts were, in contrast to many other contexts, every error can change the following punishment. For instance, stealing food multiple times may evoke exclusion from feeding or unprovoked violence in the future. Relative frequencies in particular already have been proven to play a role in decisions regarding combat (Wilson et al., 2002; Franks & Partridge, 1993 and McComb et al., 1994). In the case of horses, being able to use information about relative frequencies could for instance be helpful

when feral stallions are fighting for another stallion's band (Boyd & Keiper, 2005). It seems plausible that the ratio of mares to young stallions would determine the value of a band since the latter are potential future competitors. Statistics in the way of this study could be used when single members of a band (like samples) are abducted: "How big are my chances of getting some mares/a mare?"

On the other hand, a growing body of research suggests that discrimination of ratios might underlie the same limits as approximate discrimination of absolute quantities (Emmerton, 2001; Jacob et al., 2012 and McCrink & Wynn, 2007). This points towards a common core system for absolute and relative frequencies. Therefore, it is also possible that statistics developed early in evolution and remained present in many descendent species, too.

From these possibilities we can derive a set of species that will shed more light onto this question: social versus solitary species, species that share the common evolutionary branch of equines and primates versus species that do not, vertebrates versus invertebrates. It would be especially interesting to examine if species that have been proven to be able of absolute quantity discrimination are also capable of discriminating relative frequencies and therefore whether the currently known number systems are truly necessary for statistical reasoning.

Undeniably, there are even more, although less likely possibilities for the evolutionary trajectory. But further investigating this ability in more and more species, especially in natural contexts, will gradually give us information about ultimate and proximate causes of statistical reasoning.

Finally, this study is yet another hint that equine cognition deserves more scientific attention. More research should focus especially on domestic horses, for one thing, because domesticated animals share hundreds or thousands of years of coevolution with

humans, thus representing an especially interesting subject for research, and for another, because we have a special responsibility to study animals we live in close relationship with to assure they are treated adequately.

References

Agrillo, C., Piffer, L. & Bisazza, A. (2010). Large Number Discrimination by Mosquitofish. *PLoS ONE, 5*(12), *e15232.*

Agrillo, C., Piffer, L., Bisazza, A. & Butterworth, B. (2012). Evidence for Two Numerical Systems That Are Similar in Humans and Guppies. *PLoS ONE, 7*(2), e31923.

Boyd, L., & Keiper, R. (2005). Behavioural ecology of feral horses. *The domestic horse. The evolution, development and management of its behaviour. Cambridge University Press, Cambridge*, 55-82.

Denison, S. & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science, 13*(5), 798-803.

Eizirik, E., Murphy, W. J. & O'Brien, S. J. (2001). Molecular Dating and Biogeography of the Early Placental Mammal Radiation. *Journal of Heredity, 92*(2), 212-219.

Emmerton, J. (2001). Pigeons' discrimination of color proportion in computer-generated visual displays. *Animal Learning and Behavior, 29*(1), 21–35.

Feh, C. & De Mazières, J. (1993). Grooming at a preferred site reduces heart rate in horses. *Animal Behaviour, 46*(6), 1991-1994.

Feigenson, L., Dehaene, S. & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences, 8*(7), 307-314.

Gross, H. J. (2014), Die Geschichte vom klugen Pferd Hans. *Biologie in unserer Zeit, 44*(4), 268–270.

Gross H. J., Pahl M., Si A., Zhu H., Tautz J. & Zhang S. (2009). Number-Based Visual Generalisation in the Honeybee. *PLoS one, 4*(1), e4263.

Hanggi, E. B. (2003). Discrimination learning based on relative size concepts in horses (*Equus caballus*). *Applied Animal Behaviour Science, 83*(3), 201–213.

Hanggi, E. B. (2005). The Thinking Horse: Cognition and Perception Reviewed. *AAEP Proceedings, 51*, 246-255.

Hanus, D. & Call, J. (2007). Discrete Quantity Judgements in the Great Apes (*Pan paniscus, Pan troglodytes, Gorilla gorilla, Pongo pygmaeus*): The Effect of Presenting Whole Sets Versus Item-by-Item. *Journal of Comparative Psychology, 121*(3), 241-249.

Houpt, K. A., Law, K., Martinisi, V. (1978). Dominance hierarchies in domestic horses. *Applied Animal Ethology, 4*(3), 273-283.

Humphrey, N. K. (1976). The social function of intellect. In Bateson, P. P. G., Hinde, R. A. (Eds.), *Growing Points in Ethology* (303–317). Cambridge: Cambridge Univ. Press.

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: the brain's code for proportions. *Trends in Cognitive Sciences*, *16*(3), 157-166.

Kaufman, E. L., Lord, M. W., Reese, T. W. & Volkmann, J. (1949). The Discrimination of Visual Number. *The American Journal of Psychology, 62*(4), 498-525.

Krüger, K. & Heinze, J. (2007). Horse sense: social status of horses (*Equus caballus*) affects their likelihood of copying other horses' behavior. *Animal Cognition, 11*(3), 431-439.

Krusche, P., Uller, C., Dicke, U. (2010). Quantity discrimination in salamanders. *Journal of Experimental Biology, 213*(11)*,* 1822-1828.

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science*, *18*(8), 740-745.

Murphy, J., Waldmann, T., & Arkins, S. (2004). Sex differences in equine learning skills and visio-spatial ability. *Applied Animal Behaviour Science*, *87*(1), 119-130.

Petrazzini, M. E. M. (2014). Trained Quantity Abilities in Horses (*Equus caballus*): A Preliminary Investigation. *Behavioral Sciences 4*(3), 213–225.

Piaget, J. & Inhelder, B. (1975). *The origin of the idea of chance in children*. (Trans L. Leake, P. Burrel, & HD Fishbein). WW Norton.

Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J. et al. (2014). Apes are intuitive statisticians. *Cognition, 131*(1), 60-68.

Roth, G. & Dicke, U. (2005). Evolution of the brain and intelligence. *TRENDS in Cognitive Science, 9*(5), 250-257.

Rugani, R., Vallortigara, G., Regolin, L. (2014). From small to large: Numerical discrimination by young domestic chicks (*Gallus gallus*). *Journal of Comparative Psychology, 128*(2), 163-171.

Sigurjónsdóttir, H., van Dierendonck, M. C. & Thórhallsdóttir, A. G. (2002). Friendship Among Horses - Rank and Kinship Matter. *In Havemeyer Foundation Workshop on Horse Behavior.*

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89-96.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*(4), 410-433.

Tversky, A. & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185 (4157)*, 1124-1131.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211 (4481)*, 453-458.

Uller, C. & Lewis, J. (2009). Horses (*Equus caballus*) select the greater of two quantities in small numerical contrasts. *Animal Cognition, 12*(5), 733-738.

Vonk, J. & Beran, J. M. (2012). Bears ‘count‘ too: quantity estimation and comparison in black bears, *Ursus americanus*. *Animal Behaviour, 84*(1), 231-238.

Wolff, A., & Hausberger, M. (1996). Learning and memorisation of two different tasks in horses: the effects of age, sex and sire. *Applied Animal Behaviour Science*, *46*(3), 137-143.

Xu, F. & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences, 105*(13)*,* 5012–5015.

Appendix A

Table 4.1

Order of the experiments, gender, pregnancy and age of each horse by name, including disqualified subjects

Name	Order of experiments	Gender	Pregnancy	Age
Accolino	2, 1, 3	gelding		15
Annabelle	3, 1, 2	mare	no	20
Caruso	1, 2, 3	gelding		16
Daisy	disqualified	mare	no	2
Dancing Queen	3, 2, 1	mare	no	7
Dark Lady	disqualified	mare	yes	
Darling	2, 1, 3	mare	no	4
Dawina	disqualified	mare	no	
Daydream	2, 3, 1	mare	no	14
Dayna	1, 3, 2	mare	no	8
Delia	3, 2, 1	mare	no	17
Delina	1, 3, 2	mare	no	10
Dorina	disqualified	mare	no	2
Fatima	3, 1, 2	mare	no	25
Fräulein	disqualified	mare	no	2
Geron	3, 2, 1	stallion		8
Gig	1, 3, 2	stallion		3
Gigolo	2, 1, 3	stallion		16

Continuation

Name	Order of experiments	Gender	Pregnancy	Age
Leandro	2, 3, 1	gelding		15
Lena	3, 2, 1	mare	no	14
Linda	3, 1, 2	mare	no	7
Lisa	1, 2, 3	mare	no	5
Nemo	1, 2, 3	stallion		17
Nero	2, 3, 1	stallion		4
Nino	2, 1, 3	stallion		4
Paloma	disqualified	mare	no	
Paola	disqualified	mare	maybe	
Patty	3, 1, 2	mare	yes	6
Peggy	1, 3, 2	mare	yes	8
Perle	3, 2, 1	mare	yes	17
Pik-Ap	3, 1, 2	gelding		20
Polly	1, 2, 3	mare	no	12
Prima	1, 3, 2	mare	no	12
Primel	disqualified	mare	yes	
Sally	2, 1, 3	mare	no	10
Vincent	2, 3, 1	gelding		13
Wesper	2, 3, 1	stallion		11

Table 4.2

Breed, purpose, place of testing and kind of food used of each horse by name, including disqualified subjects (*)

Name	Breed	Purpose	Test place	Food
Accolino	Deutsches Sportpferd	lessons	box	carrot
Annabelle	Deutsches Sportpferd	put out to grass	box	carrot
Caruso	Deutsches Sportpferd	lessons	box	carrot
Daisy*	Deutsches Sportpferd	young	-	carrot
Dancing Queen	Deutsches Sportpferd	young	box	carrot
Dark Lady*	Deutsches Sportpferd	breeding	box	apple
Darling	Deutsches Sportpferd	young	box	carrot
Dawina*	Deutsches Sportpferd	leisure	box	carrot
Daydream	Deutsches Sportpferd	lessons	box	carrot
Dayna	Deutsches Sportpferd	young	box	carrot
Delia	Deutsches Sportpferd	breeding	box	carrot
Delina	Deutsches Sportpferd	young	box	carrot
Dorina*	Deutsches Sportpferd	young	-	carrot
Fatima	Sachsen-Anhaltiner	leisure	box	carrot
Fräulein*	Deutsches Sportpferd	young	box	carrot
Geron	Shetland pony	breeding	outside	carrot
Gig	Shetland pony	young	outside	apple
Gigolo	Shetland pony	breeding	outside	carrot

Continuation

Name	Breed	Purpose	Test place	Food
Leandro	Deutsches Sportpferd	lessons	box	carrot
Lena	Shetland pony	breeding	box	carrot
Linda	Shetland pony	breeding	outside	carrot
Lisa	Shetland pony	breeding	outside	carrot
Nemo	Shetland pony	breeding	outside	carrot
Nero	Shetland pony	breeding	outside	carrot
Nino	Shetland pony	young	outside	carrot
Paloma*	Shetland pony	breeding	outside	carrot
Paola*	Shetland pony	breeding	outside	carrot
Patty	Shetland pony	breeding	outside	carrot
Peggy	Shetland pony	breeding	box	carrot
Perle	Shetland pony	breeding	outside	carrot
Pik-Ap	Deutsches Sportpferd	lessons	box	carrot
Polly	Shetland pony	breeding	outside	carrot
Prima	Shetland pony	breeding	outside	carrot
Primel*	Shetland pony	breeding	outside	carrot
Sally	Schweres Warmblut	leisure	box	apple
Vincent	Schweres Warmblut	leisure	box	carrot
Wesper	Shetland pony	breeding	box	carrot

Appendix B

Detailed information about training performance of each horse

Name	Required trials first pretest	Number of sessions first pretest	Required trials second pretest[a]
Accolino	30	1	15
Annabelle	135	3	12
Caruso	40	1	33
Dancing Queen	17	1	15
Dark Lady	105	3	18
Darling	29	1	28
Daydream	68	2	34
Dayna	90	2	51
Delia	81	2	34
Delina	48	1	19
Fatima	12	1	22
Geron	75	1	22
Gig	42	1	34
Gigolo	120	2	18
Leandro	52	1	32
Lena	25	1	12
Linda	38	1	22
Lisa	58	1	24
Nemo	46	1	13

Continuation

Name	Required trials first pretest	Number of sessions first pretest	Required trials second pretest[a]
Nero	32	1	37
Nino	35	1	15
Patty	29	1	33
Peggy	86	2	43
Perle	83	2	12
Pik-Ap	26	1	12
Polly	38	1	49
Prima	138	2	14
Sally	32	2	12
Vincent	43	1	30
Wesper	115	3	45

[a] In the second pretest, all subjects completed the indicated number of trials in one session.

Appendix C

Detailed information about test performance of each horse

	Success rate			First trial[a]		
Name	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp. 1**	**Exp. 2**	**Exp. 3**
Accolino	0.83	0.42	0.58	1	0	0
Annabelle	0.67	0,75	0.42	1	1	1
Caruso	0.50	0.67	0.58	0	1	1
Dancing Queen	0,75	0.50	0.50	1	0	0
Darling	0.67	0,75	0.50	1	1	0
Daydream	1.00	0.58	0.50	1	1	1
Dayna	0.67	0,75	0.42	1	1	0
Delia	0.58	0.58	0.42	1	0	0
Delina[b]	0.50	0.50	0.58	1	0	1
Fatima	0.58	0.83	0.50	1	1	0
Geron	0,75	0.50	0.50	1	0	0
Gig	0.50	0.67	0.33	1	0	1
Gigolo	0.50	0.42	0.50	0	1	1
Leandro	0.83	0.67	0.58	1	1	1
Lena	0.67	0,75	0.50	1	1	1
Linda	0.58	0.83	0.50	1	1	0
Lisa	0.50	0.50	0.33	1	1	1
Nemo	0.50	0.50	0.50	1	0	1
Nero	0.58	0.33	0.50	1	0	1

Continuation

	Success rate			First trial[a]		
Name	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp. 1**	**Exp. 2**	**Exp. 3**
Nino	1.00	0.58	0.50	1	1	0
Patty	0.58	0.67	0.50	1	1	1
Peggy	0.50	0.67	0.50	1	1	0
Perle	0.58	0.67	0.50	0	1	1
Pik-Ap	0.58	0.67	0.58	0	1	1
Polly[c]	0.45	0.58	0.50	0	1	0
Prima	0.58	0,75	0.58	1	0	0
Sally	0,75	0.50	0.58	1	0	1
Vincent	0,75	0.50	0.58	0	1	0
Wesper	0.83	0.50	0.50	1	0	1

[a] Ones indicate a correct trial, zeros an incorrect trial.

[b] Subject Delina at first only completed 10 trials in Exp. 1 and was therefore tested again on another day. Only the second, complete session was included in the analyses.

[c] Subject Polly only completed 11 trials in Exp. 1 because one trial was accidently counted double. This was only discovered after finishing data collection, during the comparison with the video tapes. Thus, Polly's performance was calculated on the basis of 11 instead of 12 trials.

Appendix D

Inferential statistics for Exp. 3 adjusted for outliers

Test	df	Test statistic	Significance	Effect size
group performance against chance	26	$t = 1.55$	.067	Cohen's $d = 0.30$
effect gender females, males	25	$t = -1.96$	.061	Cohen's $d = -0.77$
effect gender mares, geldings, stallions	2 24	$F = 8.15$	.002	$\eta^2 = 0.40$
mares – geldings			.002	
stallions – geldings			.008	
stallions – mares			1.00	
effect of age	25		.783	$r = .056$
effect of breed	11.7	$t = 0.38$	.709	Cohen's $d = 0.16$
effect of purpose	25	$t = -2.73$	.011	Cohen's $d = -1.35$

Note. Descriptive statistics for Exp. 3 are $M = 0.52$, $SD = 0.05$.

Acknowledgements

My special thanks go to Christoph and Matthias Heinrich and Franziska Peter for giving me free access to their horses and the stables and for their cooperation during data collection. I would also like to thank Johanna Eckert from the Max Planck Institute for Evolutionary Anthropology in Leipzig for answering my questions about Rakoczy et al.'s study. Last but not least, I want to thank the Leipzig Research Center for Early Child Development for providing me with camera equipment.

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

This paper was not previously presented to another examination board and has not been published in German, English or any other language.

Jena, 06.09.2016